(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 625 757 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **94303269.8**

(51) Int. Cl.[5] : **G06F 15/401,** G06F 15/403

(22) Date of filing : **06.05.94**

(30) Priority : **07.05.93 US 60429**

(43) Date of publication of application :
23.11.94 Bulletin 94/47

(84) Designated Contracting States :
DE FR GB IT

(71) Applicant : **CANON KABUSHIKI KAISHA**
30-2, 3-chome, Shimomaruko,
Ohta-ku
Tokyo (JP)

(72) Inventor : **Devito, Jonathan**
21683 Summit Road
Los Gatos, California 95030 (US)

Inventor : **Garland, Harry**
27555 Purissima Road
Los Altos Hills, California 94022 (US)
Inventor : **Hunter, Ken**
151A Saturn Street
San Francisco, California 94114 (US)
Inventor : **May, Gerald A**
12149 Atrium Dr.
Saratoga, California 95070 (US)
Inventor : **Roberts, Michael G**
950 High School Way 3201
Mountain View, California 94041 (US)

(74) Representative : **Beresford, Keith Denis Lewis et al**
BERESFORD & Co.
2-5 Warwick Court
High Holborn
London WC1R 5DJ (GB)

(54) Selective document retrieval method and system.

(57) Method and system for storing and selectively retrieving information, such as words, from a document set. The method includes generating an image data set representative of the information contained in the document set. The method also involves generating a text data set representative of a text portion of the information contained in the document set. A text-image correspondence (TIC) table is generated that includes data representative of coordinates information corresponding to each phrase of the document set. A search phrase is identified in response to user-specified search criteria and the search phrase is identified in the text image data set. Then, the TIC table is used to identify the coordinates information corresponding to the search phrase identified in the text data set. A display of the portion of the page containing the search phrase is generated using the coordinates information.

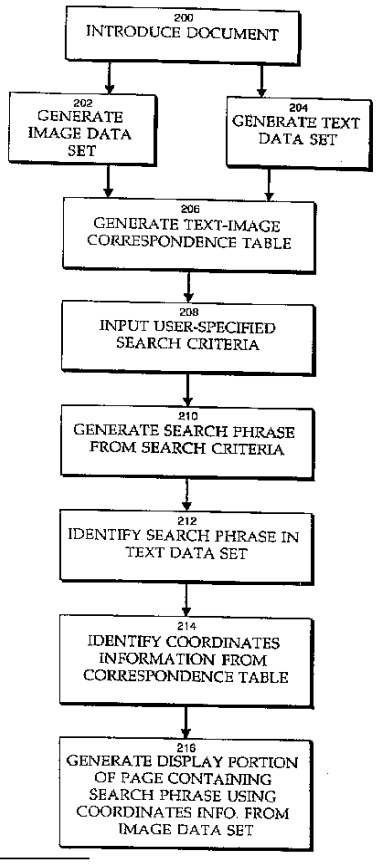


FIGURE 2

EP 0 625 757 A1

The present invention relates generally to the field of document storage and retrieval systems of the type used for multiple document manipulation. Specifically, the invention relates to a method and system for storing documents as both text and graphics, searching a document set for a search term, then generating a graphics display of the portion of a page from the document set containing the search term.

Electronic storage of documents has facilitated the handling of large volumes of documents, such as those handled by hospitals, universities, government institutions, and the like. Typically, the documents are entered into massive storage systems by use of a scanner system that converts text into electronic data. Documents primarily containing text can readily be scanned and stored in various electronic forms in this manner. It is also possible to both scan a document and perform optical character recognition on the document to create stored text and image versions of the document.

Once the documents are stored, there is a need to retrieve selected documents, or selected pages from the documents. For example, a hospital may choose to maintain all patient charts in a computer storage system, and have the ability to selectively call up a document for viewing and editing.

Typical existing systems address the need for selectively retrieving a document by assigning an index to the document as it is entered into storage. The index may be a system-generated or a user-defined code. The code then is stored together with the document. To retrieve a document, a user must enter the appropriate code associated with the desired document. Other systems use predetermined key words extracted from the document which the user may then use to subsequently retrieve a document.

The problem encountered with such systems is that a user must know the index, or code, associated with a desired document. If a user enters an inappropriate index or code, then the target document may not be retrieved.

In one presently available commercial system for document storage and retrieval, documents are stored in both text and graphics form. A user may enter a search term, which may be any term in the document and is not limited to a predetermined index term. The retrieved document may then be displayed in both text form and graphics form. However, the system retrieves an entire page of the document containing the search word. Since most screen displays do not have the capacity for full page text display, the user must search through the text on the screen display to locate the search term. This process is time consuming and inconvenient for a user, especially a user working with large volume document retrieval

Thus, there remains a need for a method and system for storing documents and selectively retrieving documents based on any user-selected search term, without requiring a pre-indexing of the documents. There also remains a need for such a method and system for retrieving the portion of the page containing the search term, which search term preferably is identified on the retrieved page.

The present invention is a method and system for storing and selectively retrieving information, such as words, from a document set using user-defined search criteria.

The method involves generating an image data set representative of the information contained in the document set. Preferably, this image data set is generated using a scanner system. The image data set is stored, preferably as a bit-map, in memory for subsequent access.

The method also involves generating a text data set representative of a text portion of the information contained in the document set. Preferably, this text data set is generated using a character recognition system that generates ASCII, Unicode, EBCDIC, JIS, or other forms of text data.

A text-image correspondence (TIC) table is generated that includes data representative of coordinates information for phrases from the image data set. A concordance table may also be generated containing concordance data associated with the information in the text data set.

A search phrase is identified, preferably in response to user-specified search criteria. The search phrase may be an exact match to a user-specified search phrase, or may be generated in accordance with some predetermined or user-defined rules. Various methods and systems for generating the search phrase may be used, and are described in further detail below.

The search phrase is identified in the text data set. Then, the TIC table is used to identify the coordinates information corresponding to the search phrase. A display of the portion of the page containing the search phrase is generated using the coordinates information. The displayed portion may contain the phrase shown in a single page, or may be multiple portions of multiple pages, each page containing a portion of the search phrase. Preferably, the search phrase is identified, e.g., by highlighting, on the display.

A system (apparatus) according to one embodiment of the invention includes a stored image data set, a stored text data set, and a text-image correspondence (TIC) table, each substantially as described above. The system includes a processor which identifies a search phrase, corresponding to a user-specified search criteria, in the text data set, and coordinates information, from the TIC table, corresponding to the search phrase. The system further includes a computer display device for displaying a portion of the page containing the search phrase, based on the identified coordinates information.

The accompanying drawings, which are incorporated in and form a part of the specification, illustrate embodiments of the present invention and, together with the description, explain the principles of the invention.

FIGURE 1 is a block diagram showing the structure of an embodiment of the invention.

FIGURE 2 is a flow chart illustrating an embodiment of the present invention.

FIGURE 3 is a flow chart illustrating an expanded portion of the flow chart of FIGURE 2.

The present invention involves a computer-implemented method and system for storing and selectively retrieving information from a document set. For purposes of the following discussion, the phrase "document set" includes a single document of one or more pages, and a combination of documents which are treated as a set for purposes of storing, grouping, and selective retrieving by a user. Thus, a document set my be of variable size, ranging from one page to thousands of pages, and typically are printed on paper, but may include information contained in or on other media. Each document set may consist of text only, graphics only, or a mixture of text and graphics.

FIGURE 1 shows a block diagram of the inventive system. The system includes a scanner device 102, including a sensor array for scanning documents and converting the image information of each document 100 to a digital information stream. The sensor array may be of any type generally commercially available. The information provided by the sensor array represents the information on the scanned document 100, and is a digital bit stream provided to a processor 104. The processor 104 may be any one of a number of suitable general purpose digital computers. The image data is stored in a storage device 106.

Alternatively, the image data may be introduced into the processor 104 via electronic media, such as facsimile or e-mail. In that system (not shown), a facsimile device or electronic mail system is used in conjunction with the present document storage and retrieval system.

Preferably, the processor 104 is provided with software for converting the image data, which may be in the form of bit-mapped information, into text data, such as ASCII character data. The text recognition software for recognizing text and generating text data from bit-mapped image data presently is available and/or can be generated from commercially available software without undue experimentation.

The recognition is accomplished by the system either by dedicated logic, by software, or by a combination of special logic and software, although software appears to be more common and more efficient in currently manufactured systems. In this context, "recognition" means the conversion of the signals representative of characters forming the image into code representations of those characters. Preferably, this would be ASCII code so that it can be imported into other programs, transmitted, or otherwise processed in conventional fashions. However, other codes may be used, such as Unicode, EBCDIC, or JIS. The invention may be used in conjunction with a variety of commercially available scanner devices.

Turning now to FIGURE 2, that shows a flow chart of an embodiment of-the inventive method as implemented in the system described above. Initially, a document is introduced 200 into an optical scanner 102 for scanning or electrooptical conversion to produce image signals representative of the information on the source document. As mentioned above, an alternative embodiment may include electronically transmitting document information to the processor 104, instead of scanning printed copies of the document set.

Following scanning of the document, or otherwise generating 202 an image data set. The image data set includes information contained on the face of the scanned source document. The image data set may be compressed, preferably using JPEG or other known compression systems. Using compression on the image data set may result in a loss of some accuracy, but is desirable to reduce the amount of required memory.

A text data set is generated 204 that corresponds to the information on the document obtained using an OCR 108 or other comparable character recognition system. FIGURE 3 is a flow chart showing an expanded version of the flow chart of FIGURE 2.

As discussed above, the information is scanned 200, or otherwise entered, to generate 202 an image data set. The image data set is then divided 304 into zones, wherein text is distinguished from graphics. Consider the following contents of an exemplary image data set.

Example:

This	is	a	document	that	has	been
scanned	into	a	computer	system	and	
segmented	into	words	using	pattern		
recognition	technology					

The division of zones is represented by the horizontal lines that appear between the lines of text. Next,

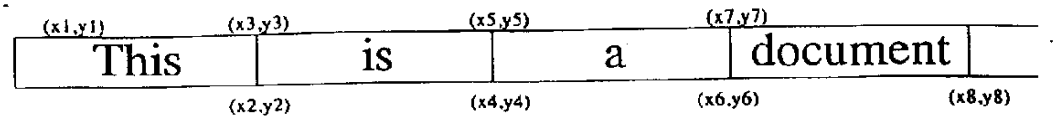
the image data set is segmented 306 by dividing the individual words. Consider now the above Example after segmenting 306:

Example:

This	is	a	document	that	has	been	
scanned	into	a	computer	system	and		
segmented	into	words	using	pattern			
recognition	technology						

The segmentation is represented by vertical lines that appear between the words of text. Using the information obtained following the zoning and segmenting steps, it is possible to identify the coordinates of the interstices. These coordinates correspond to the intersections of the lines formed by the zoning 304 and segmenting 306 steps, as shown in the Example:

Example:



A parsing also may be used to eliminate certain data, such as punctuation and formatting, to reduce the data set. The text data set is generated 204 from the stored image data set, preferably using OCR technology.

A text-image correspondence (TIC) table is generated 206 that contains the interstitial coordinates information. The TIC table includes information correlating the text data set with the coordinates information of the image data set. An exemplary TIC table is as follows:

Text data set	Image data set
This	x1,y1 x2,y2
is	x3,y3 x4,y4
a	x5,y5 x6,y6
document	x7,y7 x8,y8

The TIC table may be retained in a storage device as a look-up table, accessible by the processor 204 of the system. The TIC table links the instances of phrases contained on the document, as maintained in the text data set, with the specific locations of the phrases as contained in the image data set.

Once the TIC table and both data sets are stored, a user may input 208 search criteria to locate a phrase within the stored document. The search criteria may be predetermined criteria, presented to the user in the form of a listing or menu of options in a graphical user interface on a computer screen display, or by other limitations imposed by the specific system. In a preferred embodiment, the user specifies the search criteria.

For purposes of this discussion, the term "phrase" shall include a word, phrase, character, or character string that a user may select from a document set. Punctuation may or may not be treated as a phrase, depending on the specific requirements of the system. In a preferred embodiment, the coordinates correspond to diagonally opposing points of a rectangle defining an area in which a phrase is located in the individual document. Other schemes for identifying the location of phrases in a document set may be used, for example by identifying other geometrically defining points.

A search phrase then is generated 110 based on the search criteria. For example, in a hospital record environment, the search criteria may simply be "all records of Jones". The resulting search phrase will be "Jones". Another example is an exact search criteria, wherein the user specified "Jones", and only an exact match "Jones" will be retrieved. Similarly, the search criteria may include non-literal searches, enabling the user to specify search criteria that would result in search phrases belonging to an identified group or class. For ex-

ample, in the hospital record environment, a user may select as the search criteria "all first names beginning with the letter 'P'". A listing of search phrases will include, for example, the phrases "Paul", "Pauline", and "Peter". It may be possible to use a variety of searching schemes available to those skilled in the art.

Once the search phrase is generated from the search criteria information, the search phrase is identified 212 in the text data set. Standard searching algorithms and schemes may be used and are readily available to those skilled in the art. These searching algorithms may include the use of concordance tables, and other schemes available for expediting the search.

The coordinates information for the identified search phrase in the text data set are identified 214 from the TIC table. The coordinates information identifies the specific location of the search phrase in the document set. Preferably, once the first occurrence of the search phrase within the text data set is identified, the TIC table is accessed to identify 214 the coordinates information. However, in another embodiment, all occurrences of a search phrase may be identified 212 prior to identifying 214 the corresponding coordinates information.

Once the coordinates information is identified 214 from the TIC table, a display of the portion of the document containing the search phrase is generated 216 on a display device 110 from the image data set. The significance of generating the display from the image data set is most notable for instances when the search results are disjoint. That is, when the search phrase is located on two lines (i.e., it is in a hyphenated form) or split between two different pages (e.g., for phrases containing two or more words). It may even be possible for a search phrase to be split between several pages. In those special instances, using coordinates information for phrases enables the system to generate a display containing the entire search phrase, even though the display may contain a portion of two different pages from the document set.

Several document display modes may be implemented. In a preferred mode, a graphic image is constructed from the image data set, providing the user with an exact image of the scanned document. The display would include both textual and non-textual portions of the document.

In a preferred embodiment, the display includes some indication of the search phrase. Specifically, the search phrase may be highlighted or otherwise graphically indicated to the user on the computer display screen 210. Since the display is generated 216 from the image data set, it is possible to backlight, or otherwise indicate the search phrase on the retrieved portion of the document set.

Claims

1. A computer-implemented method for storing and selectively retrieving information contained in a document set including at least one page, comprising:
   A. generating an image data set representative of the information contained in the document set;
   B. storing the image data set in a first memory storage device;
   C. generating a text data set representative of a text portion of the information contained in the document set;
   D. storing the text data set in a second memory storage device;
   E. generating a text-image correspondence table including data representative of coordinates information corresponding to each phrase in the document set;
   F. identifying a search phrase, corresponding to a user-specified search criteria, in the text data set;
   G. identifying the coordinates information, from the text-image correspondence table, corresponding to the search phrase identified in the text data set; and
   H. generating a display of a portion of a page of the document set containing at least a portion of the search phrase, based on the identified coordinates information.

2. The method of claim 1, wherein Step H further comprises indicating at least a portion of the search phrase on the display.

3. The method of claim 1, wherein Step C comprises performing optical character recognition on the stored image data set to generate the text data set.

4. The method of claim 3, wherein the step of performing optical character recognition further comprises segmenting the text portion and the graphics portion from the information contained in the document set.

5. The method of claim 1, wherein the image data set comprises a bit-map data set.

6. The method of claim 5, wherein Step C comprises performing optical character recognition on the stored

image data set to generate the text data set.

7. The method of claim 1, wherein the text data set comprises one from the group consisting of: ASCII; Unicode; EBCDIC; and JIS.

8. The method of claim 1, wherein Step B further comprises compressing the image data set.

9. The method of claim 1, wherein Step H comprises generating a display of a portion of a plurality of pages of the document, each page containing at least a portion of the search term.

10. A computer-implemented method for storing and selectively retrieving information contained in a document set including a plurality of pages, comprising:
A. generating an image data set representative of the information contained in the document set;
B. storing the image data set in a first memory storage device;
C. generating a text data set representative of a text portion of the information contained in the document set;
D. storing the text data set in a second memory storage device;
E. generating a text-image correspondence table including data representative of coordinates information corresponding to each phrase of the document set;
F. generating a set of non-literal search terms, in accordance with a predetermined set of rules, corresponding to user-specified search criteria.
G. identifying at least one of the non-literal search terms in the text data set;
H. identifying the coordinates information, from the text-image correspondence table, corresponding to the non-literal search term identified in the text data set; and
I. generating a display of a portion of a page of the document set containing at least a portion of the non-literal search term, based on the identified coordinates information.

11. The method of claim 10, wherein Step G further comprises indicating at least a portion of the non-literal search term on the display.

12. The method of claim 10, wherein Step C comprises performing optical character recognition on the stored image data set to generate the text data set.

13. The method of claim 12, wherein the step of performing optical character recognition further comprises segmenting the text portion and the graphics portion from the information contained in the document set.

14. The method of claim 10, wherein the image data set comprises a bit-map data set.

15. The method of claim 14, wherein Step C comprises performing optical character recognition on the stored image data set to generate the text data set.

16. The method of claim 10, wherein the text data set comprises one from the group consisting of: ASCII; Unicode; EBCDIC; and JIS.

17. The method of claim 10, wherein Step B further comprises compressing the image data set.

18. The method of claim 12, wherein Step I comprises generating a display of a portion of a plurality of pages of the document set, each page containing at least a portion of the non-literal search term.

19. A computer-implemented method for storing and selectively retrieving information contained in a document set including at least one page, and including a stored image data set representative of the information contained in the document set and a stored text data set representative of a text portion of the information contained in the document set, the method comprising;
A. generating a text-image correspondence table including data representative of coordinates information corresponding to each phrase of the document set;
B. identifying a search phrase, corresponding to a user-specified search criteria, in the text data set;
C. identifying the coordinates information, from the text-image correspondence table, corresponding to the search phrase identified in the text data set; and
D. generating a display of a portion of a page of the document set containing at least a portion of the search phrase, based on the identified coordinates information.

20. A computer-implemented method for storing and selectively retrieving information contained in a document set including at least one page, and including a stored image data set representative of the information contained in the document set and a stored text data set representative of a text portion of the information contained in the document set, the method comprising:
    A. generating a text-image correspondence table including data representative of coordinates information corresponding to each phrase of the data set;
    B. generating a set of non-literal search terms, in accordance with a predetermined set of rules, corresponding to user-specified search criteria;
    C. identifying at least one of the non-literal search terms in the text data set;
    D. identifying the coordinates information, from the text-image correspondence table, corresponding to the non-literal search term identified in the text data set; and
    E. generating a display of a portion of a page of the document set containing at least a portion of the non-literal search term, based on the identified coordinates information.

21. A system for storing and selectively retrieving information contained in a document set including at least one page, comprising:
    A. a stored image data set representative of the information contained in the document set;
    B. a stored text data set representative of a text portion of the information contained in the document set;
    C. a text-image correspondence table including data representative of coordinates information corresponding to each phrase of the document set;
    D. means for identifying a search phrase, corresponding to a user-specified search criteria, in the text data set;
    E. means for identifying the coordinates information, from the text-image correspondence table, corresponding to the search phrase identified in the text data set; and
    F. means for generating a display of a portion of a page of the document set containing at least a portion of the search phrase, based on the identified coordinates information.

22. The system of claim 21, further comprising means for indicating at least a portion of the search phrase on the display.

23. The system of claim 21, further comprising an optical character recognition (OCR) device for performing OCR on the stored image data set to generate the text data set.

24. The system of claim 21, wherein the image data set comprises a bit-map data set.

25. The system of claim 21, wherein the text data set comprises one from the group consisting of: ASCII; Unicode; EBCDIC; and JIS.

26. A computer-implemented method, comprising;
    A. generating a text-image correspondence table including data representative of coordinates information corresponding to each phrase of a document set;
    B. identifying a search phrase, corresponding to a user-specified search criteria, in a stored text data set representative of a text portion of the information contained in the document set;
    C. identifying the coordinates information, from the text-image correspondence table, corresponding to the search phrase identified in the text data set; and
    D. generating a display of a portion of a page of the document set containing at least a portion of the search phrase, based on the identified coordinates information.

27. The method of claim 26, wherein Step D further comprises indicating at least a portion of the search phrase on the display.

28. The method of claim 26, wherein Step C comprises performing optical character recognition on an image data set of the document set.

29. The method of claim 28, wherein the step of performing optical character recognition further comprises segmenting the text portion and the graphics portion from the information contained in the document set.

30. The method of claim 28, wherein the image data set comprises a bit-map data set.

31. The method of claim 26, wherein the text data set comprises one from the group consisting of: ASCII; Unicode; EBCDIC; and JIS.

32. The method of claim 28, wherein Step C further comprises compressing the image data set.

33. The method of claim 26, wherein Step D comprises generating a display of a portion of a plurality of pages of the document, each page containing at least a portion of the search term.

34. An automated document retrieval apparatus wherein documents in image form are identified and retrieved from a database on the basis of a search pattern specified in text form, a part of the document image displayed or highlighted being determined automatically on the basis of the search pattern.

35. An apparatus as claimed in claim 34 wherein a text version of each stored document image can be stored and used for identifying the document to be retrieved, and correspondence data can be stored to identify corresponding parts of the document in the text and image forms.

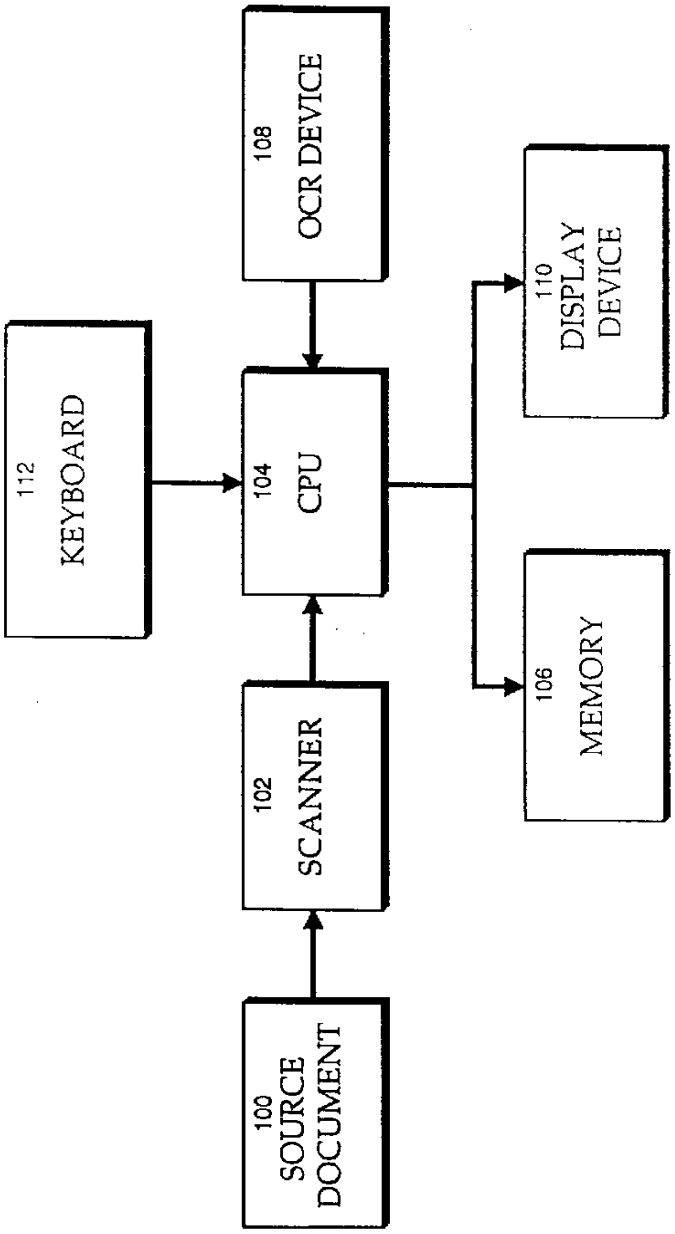

100
SOURCE
DOCUMENT
102
SCANNER
104
CPU
106
MEMORY
108
OCR DEVICE
110
DISPLAY
DEVICE
112
KEYBOARD

FIGURE 1

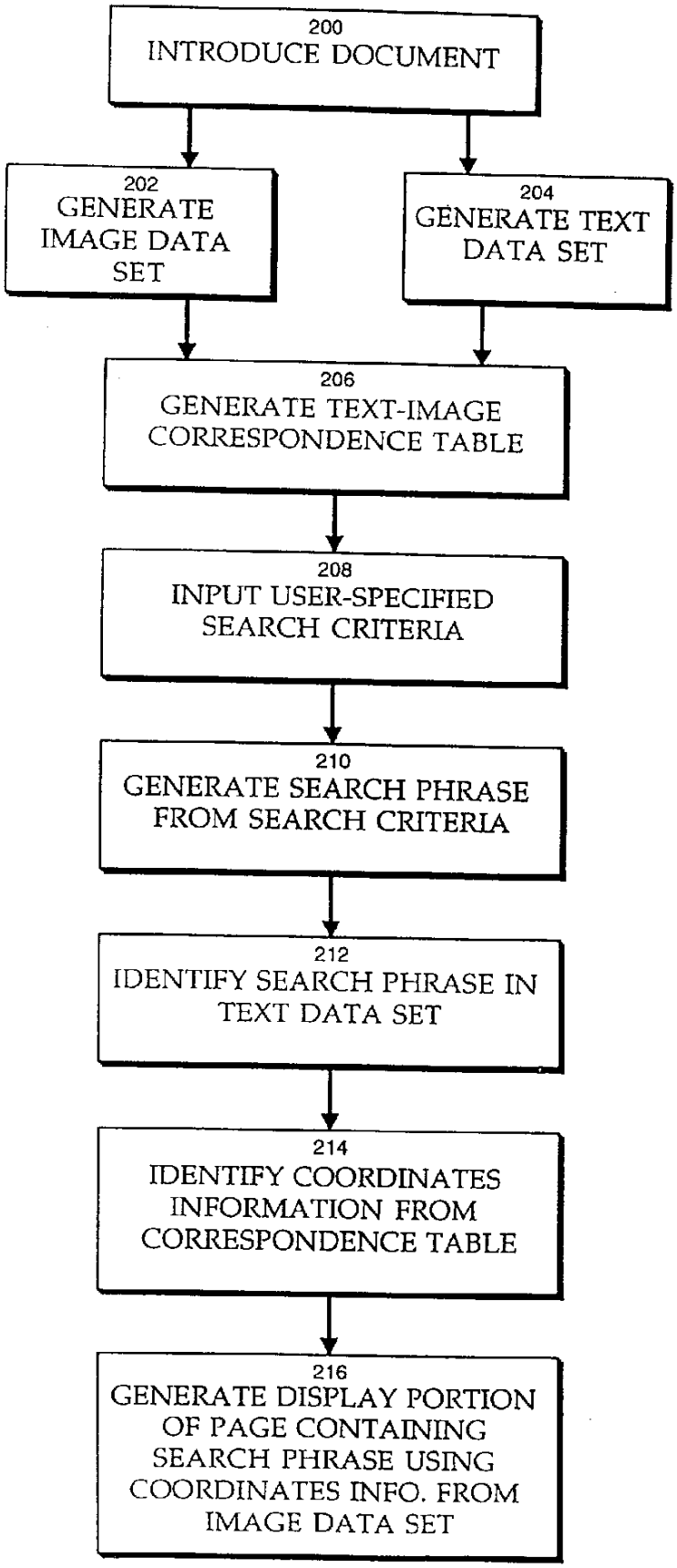

200
INTRODUCE DOCUMENT
202
GENERATE IMAGE DATA SET
204
GENERATE TEXT DATA SET
206
GENERATE TEXT-IMAGE CORRESPONDENCE TABLE
208
INPUT USER-SPECIFIED SEARCH CRITERIA
210
GENERATE SEARCH PHRASE FROM SEARCH CRITERIA
212
IDENTIFY SEARCH PHRASE IN TEXT DATA SET
214
IDENTIFY COORDINATES INFORMATION FROM CORRESPONDENCE TABLE
216
GENERATE DISPLAY PORTION OF PAGE CONTAINING SEARCH PHRASE USING COORDINATES INFO. FROM IMAGE DATA SET

FIGURE 2

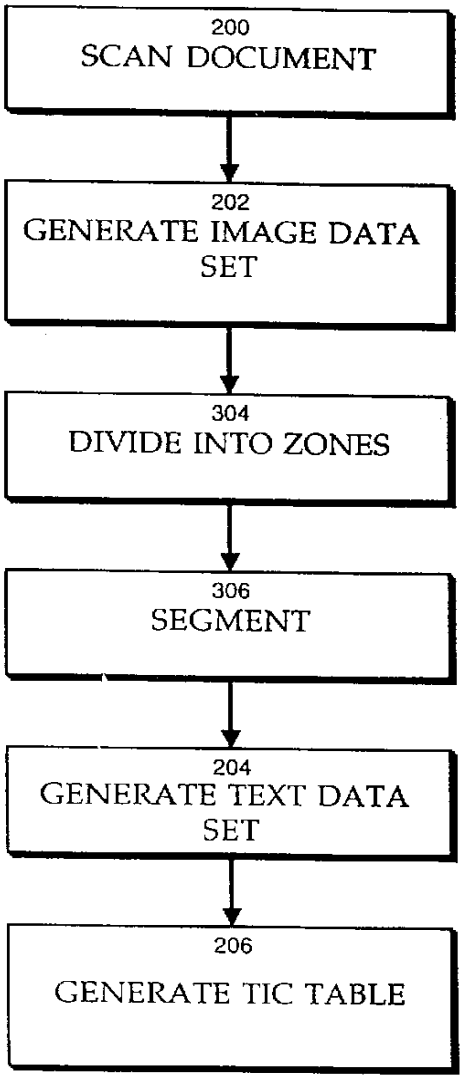

200
SCAN DOCUMENT
202
GENERATE IMAGE DATA SET
304
DIVIDE INTO ZONES
306
SEGMENT
204
GENERATE TEXT DATA SET
206
GENERATE TIC TABLE

FIGURE 3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 94 30 3269

DOCUMENTS CONSIDERED TO BE RELEVANT

	Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.5)
1	E	EP-A-0 596 247 (MOTOROLA, INC) 11 May 1994 * page 2, line 34 - line 53 *	1-35	G06F15/401 G06F15/403
1	X	EP-A-0 424 803 (FRÖSSL, HORST) 2 May 1991 * abstract * * column 6, line 42 - column 10, line 19; claims *	1-35	
5	A	PATENT ABSTRACTS OF JAPAN vol. 16, no. 68 (P-1314) 19 February 1992 & JP-A-32 060 768 (FUJI ELECTRIC CO.) 20 November 1991 * abstract *	1-35	
3	A	EP-A-0 465 818 (FRÖSSL, HORST) 15 January 1992 * abstract; claims 1-3 *	1-35	

TECHNICAL FIELDS SEARCHED (Int.Cl.5)

G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	10 August 1994	Fournier, C

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)